INSTITUTE
FOR LAW & AI

The Governance Misspecification Problem

LawAI Working Paper Series, No. 3-2024

Christoph Winter, Charlie Bullock

October 2024

law-ai.org

LawAI Working Paper No. 3-2024
Last update: October 2024
Comments welcome: charlie.bullock@law-ai.org

The Governance Misspecification Problem

Christoph Winter[*] and Charlie Bullock[†]

Abstract

Legal rules promulgated to govern emerging technologies often rely on proxy terms and metrics in order to indirectly effectuate background purposes. A common failure mode for this kind of rule occurs when, due to incautious drafting or unforeseen technological developments, a proxy ceases to function as intended and renders a rule ineffective or counterproductive. Borrowing a concept from the technical AI safety literature, we call this phenomenon the "governance misspecification problem." This article draws on existing legal-philosophical discussions of the nature of rules to define governance misspecification, presents several historical case studies to demonstrate how and why rules become misspecified, and suggests best practices for designing legal rules to avoid misspecification or mitigate its negative effects. Additionally, we examine a few proxy terms used in existing AI governance regulations, such as "frontier AI" and "compute thresholds," and discuss the significance of the problem of misspecification in the AI governance context.

Keywords

AI safety; governance misspecification; Hart; Fuller; purposivism; proxy terms

[*] Harvard University, Cambridge, MA, USA / Institute for Law & AI, Cambridge, MA, USA. Email: christoph_winter@fas.harvard.edu.
[†] Institute for Law & AI, Cambridge, MA, USA. Email: charlie.bullock@law-ai.org.

In technical Artificial Intelligence ("AI") safety research, the term "specification" refers to the problem of defining the purpose of an AI system so that the system behaves in accordance with the true wishes of its designer.[1] Technical researchers have suggested three categories of specification: "ideal specification," "design specification," and "revealed specification."[2] The ideal specification, in this framework, is a hypothetical specification that would create an AI system completely and perfectly aligned with the desires of its creators. The design specification is the specification that is actually used to build a given AI system. The revealed specification is the specification that best describes the actual behavior of the completed AI system. "Misspecification" occurs whenever the revealed specification of an AI system diverges from the ideal specification—i.e., when an AI system does not perform in accordance with the intentions of its creators.

The fundamental problem of specification is that "it is often difficult or infeasible to capture exactly what we want an agent to do, and as a result we frequently end up using imperfect but easily measured proxies."[3] Thus, in a famous example from 2016, researchers at OpenAI attempted to train a reinforcement learning agent to play the boat-racing video game CoastRunners, the goal of which is to finish a race quickly and ahead of other players.[4] Instead of basing the AI agent's reward function on how it placed in the race, however, the researchers used a proxy goal that was easier to implement and rewarded the agent for maximizing the number of points it scored. The researchers mistakenly assumed that the agent would pursue this proxy goal by trying to complete the course quickly. Instead, the AI discovered that it could achieve a much higher score by refusing to complete the course and instead driving in tight circles in such a way as to repeatedly collect a series of power-ups while crashing into other boats and occasionally catching on fire.[5] In other words, the design specification ("collect as many points as possible") did not correspond well to the ideal specification ("win the race"), leading to a disastrous and unexpected revealed specification (crashing repeatedly and failing to finish the race).

This article applies the misspecification framework to the problem of AI governance. The resulting concept, which we call the "governance misspecification problem," can be briefly defined as occurring when a legal rule relies unsuccessfully on proxy terms or metrics. By framing this new concept in terms borrowed from the technical AI safety literature, we hope to incorporate valuable insights from that field into legal-philosophical discussions around the nature of rules and, importantly, to help technical researchers understand the philosophical and policymaking challenges that AI governance legislation and regulation poses.

---

[1] Ortega et al. (2018).
[2] Id.
[3] Clark & Amodei (2016).
[4] Id.
[5] Id.

It is generally accepted among legal theorists that at least some legal rules can be said to have a purpose or purposes and that these purposes should inform the interpretation of textually ambiguous rules.[6] The least ambitious version of this claim is simply an acknowledgment of the fact that statutes often contain a discrete textual provision entitled "Purpose," which is intended to inform the interpretation and enforcement of the statute's substantive provisions.[7] More controversially, some commentators have argued that all or many legal rules have, or should be constructively understood as having, an underlying "true purpose," which may or may not be fully discoverable and articulable.[8]

The purpose of a legal rule is analogous to the "ideal specification" discussed in the technical AI safety literature. Like the ideal specification of an AI system, a rule's purpose may be difficult or impossible to perfectly articulate or operationalize, and rulemakers may choose to rely on a legal regime that incorporates "imperfect but easily measured proxies"—essentially, a design specification. "Governance misspecification" occurs when the real-world effects of the *legal regime* (analogous to the design specification) *as interpreted and enforced* (analogous to the revealed specification) fail to effectuate the *rule's intended purpose* (analogous to the ideal specification).

Consider the hypothetical legal rule prohibiting the presence of "vehicles" in a public park, famously described by the legal philosopher H.L.A. Hart.[9] The term "vehicles," in this rule, is presumably a proxy term intended to serve some ulterior purpose,[10] although fully discovering and articulating that purpose may be infeasible. For example, the rule might be intended to ensure the safety of pedestrians in the park, or to safeguard the health of park visitors by improving the park's air quality, or to improve the park's atmosphere by preventing excessive noise levels. More realistically, the purpose of

---

[6] See Manning (2006) (discussing contemporary and historical distinctions between the interpretive doctrines of textualism and purposivism and acknowledging that "because textualists understand that speakers use language purposively, they recognize that evidence of purpose… may also form an appropriate ingredient of the context used to define the text."); but see Easterbrook (1983: 536) ("The philosophy of language, and most particularly the work of Ludwig Wittgenstein, has established that sets of words do not possess intrinsic meanings and cannot be given them; to make matters worse, speakers do not even have determinative intents about the meanings of their own words.") (citing Wittgenstein (1953: §§ 138–242)); see also Eskridge (1990). Purposivism is to be distinguished from intentionalism in that purposivism does not necessarily require belief in or assign importance to the subjective intent of a legislative body, a concept which has been convincingly critiqued as incoherent. See Radin (1930); Easterbrook (1983, 1990). Instead, purposivists may rely on, e.g., a constructive "objectified" legislative intent, based on the (possibly counterfactual) assumption "that the legislature was made up of reasonable persons pursuing reasonable purposes reasonably." Hart & Sacks (1995: 1378). One recent empirical study found that more than 77% of law professors endorsed a purposivist approach to statutory interpretation, while 60% endorsed textualism and 54% endorsed intentionalism. Martínez & Tobia (2023: 48).

[7] See, e.g., Nuclear Energy Innovation and Modernization Act, Pub. L. No. 115-439, § 2, 132 Stat. 5565, 5565 (2019); see generally Stack (2019).

[8] Fuller (1956: 700), (1958).

[9] Hart (1958: 607).

[10] See Manning (2011: 115); Radin (1930: 876).

the rule might be some complex weighted combination of all of these and numerous other more or less important goals. Whether the rule is misspecified depends on whether the rule's purpose, whatever it is, is furthered by the use of the proxy term "vehicle."

Hart used the "no vehicles in the park" rule in an attempt to show that the word "vehicle" had a core of concrete and settled linguistic meaning (an automobile is a vehicle) as well as a semantic "penumbra" containing more or less debatable cases such as bicycles, roller skates, toy cars, and airplanes. The rule, in other words, is textually ambiguous, although this does not necessarily mean that it is misspecified.[11] Because the rule is ambiguous, a series of difficult interpretive decisions may have to be made regarding whether a given item is or is not a vehicle. At least some of these decisions, and the costs associated with them, could have been avoided if the rulemaker had chosen to use a more detailed formulation in lieu of the term "vehicle,"[12] or if the rulemaker had issued a statement clarifying the purpose of the rule.[13]

Although the concept of misspecification is generally applicable to legal rules, misspecification tends to occur particularly frequently and with serious consequences in the context of laws and regulations governing poorly-understood emerging technologies such as artificial intelligence. Again, consider "no vehicles in the park." Many legal rules, once established, persist indefinitely even as the technology they govern changes fundamentally.[14] The objects to which the proxy term "vehicle" can be applied will change over time; electric wheelchairs, for example, may not have existed when the rule was originally drafted, and airborne drones may not have been common. The introduction

[11] The effect of an ambiguous rule is to delay the time at which the rule's meaning is determined, and to delegate the task of determining said meaning to the parties charged with interpreting or enforcing the rule. There are potential benefits to this approach. For instance, new information may become available in the interval between enactment and interpretation, or the parties charged with enforcing the rule may be better equipped than the rule's drafter to calibrate the rule's scope. But ambiguity comes with potential costs as well, such as uncertainty about how the rule will be applied and actual or perceived unfairness if the rule is applied in an unexpected manner. See generally Maggs (1992); Schauer (1988: 539–542); Katzmann (2012).

[12] Whether or not it is ever possible to completely eliminate the possibility of penumbral edge-cases is beside the point. Reducing the size of the set of all plausible interpretive questions may reduce the number of questions that actually arise, even if the set's size is reduced only from a greater to a lesser infinity. A law that definitively addresses whether roller skates, bicycles, toy cars, and airplanes are vehicles may still fail to account for electric wheelchairs, but since a finite number of plausible potential vehicles will actually enter or be prevented from entering the park in a given year, a more detailed law may significantly reduce the number of difficult questions that are actually presented for adjudication.

[13] Hart and his opponent Lon Fuller agreed that the purpose of the "no vehicles in the park" rule was relevant to deciding the "penumbral" cases. See Hart (1958: 614). They disagreed only as to whether there also existed a separate category of "core" cases in which no inquiry into purpose was necessary. Fuller argued, using the example of a World War II Jeep mounted on a pedestal as a park monument, that even the examples identified by Hart as part of the "core" should be examined in light of the rule's purpose. Fuller (1958: 663).

[14] See Freitas-Groff (2023) (showing that even policies passed by slim majorities tend to persist for "puzzlingly long periods of time").

of these new potential "vehicles" is extremely difficult to account for in an original design specification.[15]

The governance misspecification problem is particularly relevant to the governance of cutting-edge AI systems. Unlike most other emerging technologies, current frontier AI systems are, in key respects, not only poorly understood but fundamentally uninterpretable by existing methods.[16] This problem of interpretability is a major focus area for technical AI safety researchers.[17] The widespread use of proxy terms and metrics in existing AI governance policies and proposals is, therefore, a cause for concern.[18]

In Section I, this article draws on existing legal-philosophical discussions of the nature of rules to further explain the problem of governance misspecification and situates the concept in the existing public policy literature. Sections II and III make the case for the importance of the problem by presenting a series of case studies to show that rules aimed at governing emerging technologies are often misspecified and that misspecified rules can cause serious problems for the regulatory regime they contribute to, for courts, and for society generally. Section IV offers a few suggestions for reducing the risk of and mitigating the harm from misspecified rules, including eschewing or minimizing the use of proxy terms, rapidly updating and frequently reviewing the effectiveness of regulations, and including specific and clear statements of the purpose of a legal rule in the text of the rule. Section V applies the conclusions of the previous Sections prospectively to several specific challenges in the field of AI governance, including the use of compute thresholds, semiconductor export controls, and the problem of defining "frontier" AI systems. Section VI concludes.

I. The Governance Misspecification Problem in Legal Philosophy and Public Policy

A number of publications in the field of legal philosophy have discussed the nature of legal rules and arrived at conclusions helpful to fleshing out the contours of the governance misspecification problem.[19] Notably, Schauer (1991) suggests the useful concepts of over- and under-inclusiveness, which can be understood as two common ways in which legal rules can become misspecified.[20] Overinclusive rules prohibit or prescribe

[15] For example, in McBoyle v. United States, 283 U.S. 25 (1931), a man who had been convicted of violating the National Motor Vehicle Theft Act (enacted in 1919, just 16 years after the invention of the airplane) for transporting a stolen airplane across state lines successfully petitioned to have his conviction overturned on the grounds that an airplane was not a "motor vehicle." The court reasoned that "in everyday speech 'vehicle' calls up the picture of a thing moving on land." Id. at 26.

[16] See Doshi-Velez & Kim (2017); Rudin et al. (2022).

[17] See, e.g., Bricken et al. (2023).

[18] See Section V, infra.

[19] See, e.g., Schauer (1991), Struchiner et al. (2020), Hadfield-Menell, D., & Hadfield, G. (2018).

[20] See Schauer (1991: 31) (discussing and defining under- and overinclusiveness).

actions that an ideally specified rule would not apply to, while underinclusive rules fail to prohibit or prescribe actions that an ideally specified rule would apply to. So, in Hart's "no vehicles in the park" hypothetical, suppose that the sole purpose of the rule was to prevent park visitors from being sickened by diesel fumes. If this were the case, the rule would be overinclusive, because it would prohibit many vehicles that do not emit diesel fumes to no end. Prohibitions or prescriptions that do not further the purpose of a rule in any way often come with costs, such as political backlash from the unnecessarily regulated parties.[21] If, on the other hand, the purpose of the rule was to prevent music from being played loudly in the park on speakers, the rule would be underinclusive, as it fails to prohibit a wide range of speakers that are not installed in a vehicle.

Ideal specification is rarely feasible, and practical considerations may dictate that a well-specified rule should rely on proxy terms that are under- or overinclusive to some extent. As Schauer (1991) explains, "Speed Limit 55" is a much easier rule to follow and enforce consistently than "drive safely," despite the fact that the purpose of the speed limit is to promote safe driving and despite the fact that some safe driving can occur at speeds above 55 miles per hour and some dangerous driving can occur at speeds below 55 miles per hour.[22] In other words, the benefits of creating a simple and easily followed and enforced rule outweigh the costs of over- and under-inclusiveness in many cases.[23]

In the public policy literature, the existing concept that bears the closest similarity to governance misspecification is "policy design fit."[24] Policy design is currently understood as including a mix of interrelated policy goals and the instruments through which those goals are accomplished, including legal, financial, and communicative mechanisms.[25] A close fit between policy goals and the means used to accomplish those goals has been shown to increase the effectiveness of policies.[26] The governance misspecification problem can be understood as a particular species of failure of policy design fit—a failure of congruence between a policy goal and a proxy term in the legal rule which is the means used to further that goal.[27]

[21] See, e.g., Vogel (2008).
[22] See Schauer (1991: 54).
[23] See Schauer (1988); Gigerenzer & Todd (1999); Gigerenzer & Engel (2006).
[24] See van Geet et al. (2021).
[25] Howlett (2009).
[26] Van Geet et al. (2021).
[27] Howlett & Rayner (2018) discuss means-ends congruence as one criterion that determines policy design effectiveness, along with "coherence" and "consistency."

II. Legal Rules Governing Emerging Technologies Are Often Misspecified

Misspecification occurs frequently in both domestic and international law and in both reactive and anticipatory regulations directed towards the regulation of new technologies. In order to illustrate how misspecification happens, and to give a sense of the significance of the governance misspecification problem in legal rules addressing emerging technologies, this Section discusses three historical examples of the phenomenon in the contexts of cyberlaw, copyright law, and nuclear anti-proliferation treaties.

Consider the Digital Millennium Copyright Act of 1998 (DMCA). Section 1201(a)(2) of the DMCA prohibits the distribution of any "technology, product, service, device, component, or part thereof" primarily designed to decrypt copyrighted material.[28] Congressman Howard Coble, one of the architects of the DMCA, stated that this provision was "drafted carefully to target 'black boxes'"—physical devices with "virtually no legitimate uses," useful only for facilitating piracy.[29] This provision was anticipatory in nature; the use of "black boxes" for the decryption of digital works was not widespread in 1998, but the drafters of the DMCA predicted that such devices would soon become an issue. In 1998, this prediction seemed a safe bet, as previous forms of piracy decryption had relied on specialized tools—the phrase "black box" is a reference to one such tool, also known as a "descrambler" and used to decrypt premium cable television channels.[30]

The feared black boxes never arrived. Instead, pirates relied on software, using decryption programs distributed for free online to circumvent anti-piracy encryptions.[31] Courts found the distribution of such programs, and even the posting of hyperlinks leading to websites containing such programs, to be violations of the DMCA.[32] In light of earlier cases holding that computer code was a form of expression entitled to First Amendment protection, this interpretation placed the DMCA into tension with the First Amendment.[33] This tension was ultimately resolved in favor of the DMCA, and the distribution of decryption programs used for piracy was prohibited.[34]

No one in Congress anticipated that the statute which had been "carefully drafted to target 'black boxes'" would be used to prohibit the distribution of lines of computer

---

[28] 17 U.S.C. § 1201(a)(2).

[29] Section-by-Section Analysis of H.R. 2281 as Passed by the United States House of Representatives on August 4, 1998, 105th Cong., at 8–9 (1998).

[30] See Ginsburg (1999: 144).

[31] See Crootof & Ard (2021: 362–63).

[32] See *Universal City Studios, Inc. v. Corley*, 273 F.3d 429 (2d Cir. 2001)

[33] See Crootof & Ard (2021: 362–63); *Bernstein v. United States*, 176 F.3d 1132, 1141 (9th Cir. 1999).

[34] See Crootof & Ard (2021: 362–63).

code, or that this would raise serious concerns regarding freedom of speech. Section 1201(a)(2), in other words, was misspecified; by prohibiting the distribution of any "technology" or "service" designed for piracy, as well as any "device," the framers of the DMCA banned more than they intended to ban and created unforeseen constitutional issues.

Misspecification also occurs in international law. The Treaty of Principles Governing the Activities of States in the Exploration and Use of Outer Space, which the United States and the Soviet Union entered into in 1967, obligated the parties "not to place in orbit around the Earth any objects carrying nuclear weapons or any other kinds of weapons of mass destruction, install such weapons on celestial bodies, or station such weapons in outer space in any other manner."[35] Shortly after the treaty was entered into, however, it became clear that the Soviet Union planned to take advantage of a loophole in the misspecified prohibition. The Fractional Orbital Bombardment System (FOBS) placed missiles into orbital trajectories around the earth, but then redirected them to strike a target on the earth's surface before they completed a full orbit.[36] An object is not "in orbit" until it has circled the earth at least once; therefore, FOBS did not violate the 1967 Treaty, despite the fact that it allowed the Soviet Union to strike at the U.S. from space and thereby evade detection by the U.S.'s Ballistic Missile Early Warning System.[37] The U.S. eventually neutralized this advantage by expanding the coverage and capabilities of early warning systems so that FOBS missiles could be detected and tracked, and in 1979 the Soviets agreed to a better-specified ban which prohibited "fractional orbital missiles" as well as other space-based weapons.[38] Still, the U.S.'s agreement to use the underinclusive proxy term "in orbit" allowed the Soviet Union to temporarily gain a potentially significant first-strike advantage.

The previous examples show the difficulty of avoiding misspecification when crafting anticipatory regulations. However, misspecification also occurs in laws and regulations directed towards existing and well-understood technologies. Take, for example, the Computer Fraud and Abuse Act (CFAA), 18 U.S.C. § 1030, which has been called "the worst law in technology."[39] The CFAA was originally enacted in 1984, but has since been amended several times, most recently in 2020.[40] Among other provisions, the CFAA criminalizes "intentionally access[ing] a computer without authorization or exceed[ing] authorized access, and thereby obtain[ing]... information from any protected

---

[35] Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, Including the Moon and Other Celestial Bodies (1967: Article IV).

[36] See Garthoff (1981: 38); Maas (2020: 203).

[37] Garthoff (1981: 38).

[38] Maas (2020: 203).

[39] See Wu, 2013.

[40] Technically, only the amendments passed in 1986 are the "Computer Fraud and Abuse Act," or CFAA, but in practice "courts and commentators use both labels to refer to the entire federal unauthorized access statute, 18 U.S.C. § 1030." Kerr, 2003, at 1598 n.11.

computer."[41] The currently operative language for this provision was introduced in 1996,[42] by which point the computer was hardly an emerging technology, and slightly modified in 2008.[43]

Read literally, the CFAA's prohibition on unauthorized access criminalizes both (a) violating a website's terms of service while using the internet, and (b) using an employer's computer or network for personal reasons, in violation of company policy.[44] In other words, a literal reading of the CFAA would mean that hundreds of millions of Americans commit crimes every week by, e.g., sharing a password with a significant other or accessing social media at work.[45] Court decisions eventually established narrower definitions of the key statutory terms ("without authorization" and "exceeds authorized access"),[46] but not before multiple defendants were prosecuted for violating the CFAA by failing to comply a website's terms of service[47] or accessing an employer's network for personal reasons in violation of workplace rules.[48]

Critics of the CFAA have discussed its flaws in terms of the constitutional law doctrines of "vagueness"[49] and "overbreadth."[50] These flaws can also be conceptualized in terms of misspecification. The phrases "intentionally accesses without authorization" and "exceeds authorized access," and the associated statutory definitions, are poor proxies for the range of behavior that an ideally specified version of the CFAA would have criminalized. The proxies criminalize a great deal of conduct that none of the stakeholders who drafted, advocated for, or voted to enact the law wanted to criminalize[51] and created

---

[41] 18 U.S.C. § 1030(a)(2). The term "protected computer" is defined in another section of the statute in a way that includes any computer, smart phone, or similar device capable of connecting to the internet. See *Van Buren v. United States*, 141 S. Ct. 1648, 1652 (2021). The phrase "exceeds authorized access" is defined at § 1030(e)(6) to mean "to access a computer with authorization and to use such access to obtain or alter information in the computer that the accesser is not entitled so to obtain or alter."

[42] The phrase "intentionally accesses a computer without authorization or exceeds authorized access" was present § 1030(a)(2) beginning in 1986, but prior to 1996 the provision was much narrower in scope, applying only when the defendant accessed certain kinds of financial records. Compare Pub. L. 99-474, 100 Stat. 1213 (Oct. 16, 1986), and Pub.L. 104-294, 110 Stat. 3491, 3508 (Oct. 11, 1996).

[43] The 2008 amendments expanded the provision's scope slightly, including by instituting the current, extremely expansive definition of "protected computer." See Kerr, 2010, at 1569–70.

[44] See Kerr, 2003; Kerr, 2010.

[45] See Curtiss (2016: 1841); Kerr (2010: 1578–87).

[46] See *Van Buren v. United States*, 141 S. Ct. 1648 (2021); *United States v. Nosal*, 676 F.3d 854 (9th Cir. 2012).

[47] See *United States v. Drew*, 259 F.R.D. 449 (C.D. Cal. 2009).

[48] See *United States v. Nosal*, 676 F.3d 854 (9th Cir. 2012).

[49] See, e.g., Kerr (2010).

[50] See, e.g., Chung (2010).

[51] This is not to say that none of the stakeholders *intended* to criminalize such behavior. Some stakeholders may have been aware that the new language was overinclusive (without realizing the full extent of its overinclusivity) and nevertheless supported it because they preferred a definitely overinclusive statute to a potentially underinclusive one. The Department of Justice issued a report analyzing the 1996 amendments which stated that, while "[c]ertainly not all computer misuse warrants federal criminal sanctions," the lack of an adequate "litmus test" for legislatively distinguishing between sanction-worthy and non-sanction-worthy misuse meant that any legislation would necessarily be either "under- or

substantial legal and political backlash against the law. This backlash led to a series of losses for federal prosecutors as courts rejected their broad proposed interpretations of the key proxy terms because, as the Ninth Circuit Court of Appeals put it, "ubiquitous, seldom-prosecuted crimes invite arbitrary and discriminatory enforcement."[52] The issues caused by poorly selected proxy terms in the CFAA, the Outer Space Treaty, and the DMCA demonstrate that important legal rules drafted for the regulation of emerging technologies are prone to misspecification, in both domestic and international law contexts and for both anticipatory and reactive rules. These case studies were chosen because they are representative of how legal rules become misspecified; if space allowed, numerous additional examples of misspecified rules directed towards new technologies could be offered.[53]

III. Consequences of Misspecification in the Regulation of Emerging Technologies

The case studies examined in the previous Section established that legal rules are often misspecified and illustrated the manner in which the problem of governance misspecification typically arises. This Section attempts to show that misspecification can cause serious issues when it occurs for both for the regulatory regime that the misspecified rule is part of and for society writ large. Three potential consequences of misspecification are discussed and illustrated with historical examples involving the regulation of emerging technologies.

A. Underinclusive Rules Can Create Exploitable Gaps in a Regulatory Regime

When misspecification results in an underinclusive rule, exploitable gaps can arise in a regulatory regime. The Outer Space Treaty of 1967, discussed above, is one example

over-inclusive." See Computer Crime and Intellectual Property Section, U.S. Department of Justice, 1998. In the fervently tough-on-crime political climate of the mid-90s, overinclusivity was seen as by far the lesser evil. It was believed that; prosecutors could be trusted to exercise their unlimited discretion wisely, prosecuting the bad actors while leaving the less culpable alone. See Pickett (2019).

[52] *United States v. Nosal*, 676 F.3d 854, 860 (9th Cir. 2012); see also *Van Buren v. United States*, 141 S. Ct. 1648 (2021); *United States v. Drew*, 259 F.R.D. 449 (C.D. Cal. 2009).

[53] See, e.g., U.S. v. Marshall, 908 F.2d 1312 (1990), discussing a criminal law used the weight of "a mixture or substance containing a detectable amount" of LSD as a proxy for the amount of LSD sold. This led to potentially absurd results; a single dose of LSD dissolved in a glass of orange juice would be punished more heavily than 10,000 doses on blotter paper, and dealers who used heavier blotter paper were punished more severely than dealers who sold the same amount of LSD on lighter paper. See also Maas (2020) at 203 (discussing New START, a bilateral nuclear arms control treaty signed in 2010 that used controls on the number of deployed nuclear warheads as a proxy for nuclear stability in order to avoid costly arms races. Subsequently, upgrades to U.S. warheads effectively tripled the lethality of U.S. nuclear submarines, achieving a strategically destabilizing result without adding any additional warheads).

of this phenomenon. Another example, which demonstrates how completely the use of a misspecified proxy term can defeat the effectiveness of a law, is the Audio Home Recording Act of 1992.[54] That statute was designed to regulate home taping, i.e., the creation by consumers of analog or digital copies of musical recordings. The legal status of home taping had been a matter of debate for years, with record companies taking the position that it was illegal and with the manufacturers of taping hardware arguing the opposite position.[55] Congress attempted to resolve the debate by creating a safe harbor for home taping that allowed for the creation of any number of analog or digital copies of a piece of music, with the caveat that royalties would have to be paid as part of the purchase price of any equipment or media used in the creation of digital copies.[56]

Congress designed the AHRA under the assumption that digital audio tape recorders (DATs) were the wave of the future and would shortly become a ubiquitous home audio appliance.[57] The statute lays out, in painstaking detail, a complex regulatory framework governing "digital audio recording devices," which phrase the statute defines to require the capability to create reproductions of "digital musical recordings."[58] "Digital audio recording devices" are required to conform to a "Serial Copy Management System" that would track the creation of digital copies, convey copyright information regarding the copied recordings, and require royalty payments.[59] Bizarrely, however, the AHRA explicitly provides that the term "digital musical recording" does not encompass recordings stored on any object "in which one or more computer programs are fixed"—i.e., computer hard drives.[60] In 1992, computers were not commonly used for listening to music, and Congress apparently failed to anticipate that computers might play a role in the creation of digital copies of musical recordings.

Of course, the DAT did not become a staple of the American household. And when the RIAA tried to sue the manufacturer of the "Rio," an early mp3 player, for failing to comply with the requirements the AHRA placed on digital audio recording devices, the Ninth Circuit found that the device was not subject to the AHRA.[61] Because the Rio was designed solely to download mp3 files from a computer hard drive, it was not capable of copying "digital musical recordings" under the AHRA's underinclusive definition of that phrase.[62] The court noted that this would "effectively eviscerate the Act," because "[a]ny recording device could evade [...] regulation simply by passing the music through a computer and ensuring that the MP3 file resided momentarily on the hard

[54] 17 U.S.C. § 1001 *et seq*.
[55] Nimmer (2004: 1332).
[56] Id. at 1333.
[57] Id.
[58] 17 U.S.C. § 1001
[59] Id. at §§ 1002–1004.
[60] Id. at § 1001(5)(B).
[61] *Recording Indus. Ass'n of Am. v. Diamond Multimedia Sys., Inc.*, 180 F.3d 1072 (9th Cir. 1999).
[62] Id. at 1075–78.

drive," but nevertheless rejected the creative reinterpretations of the AHRA offered by the music industry as contrary to the plain language of the statute.[63] As a result, the AHRA was effectively a dead letter less than six years after being enacted.[64]

The most obvious lesson of the utter failure of the AHRA is that Congress acted with insufficient epistemic humility by creating legislation confidently designed to address one specific technology that had not, at the time of legislation, been adopted by any significant portion of the population. But this failure of humility manifested as a failure of specification. The purpose of the statute, as articulated in a Senate report, included the introduction of a "serial copy management system that would prohibit the digital serial copying of copyrighted music" and a "royalty payment system that provides modest compensation to the various elements of the music industry for the digital home recordings of copyrighted music."[65] By crafting a law that applied only to "digital audio recording devices" and defining that proxy term in a way that failed to account for more than one possible path for technological development, Congress completely failed to accomplish those purposes. If the proxy in question had not been constructed in such a way as to exclude any recording acquired or passed through a computer, the Rio and eventually the iPod might well have fallen under the AHRA's royalty scheme, and music copyright law in the U.S. would have developed down a course more consistent with the ideal specification of the AHRA.

B. Overinclusive Rules Can Create Pushback and Enforcement Challenges

Misspecification can also create overinclusive rules, like the Computer Fraud and Abuse Act and § 1201(a)(2) of the Digital Millennium Copyright Act, discussed above in Section II. As those examples showed, overinclusive rules may give rise to legal and political challenges, difficulties with enforcement, and other unintended and undesirable results. These effects can, in some cases, be so severe that they require a total repeal of the rule in question.

This was the case with a 2011 Nevada statute authorizing and regulating driverless cars. AB511, which was the first law of its kind enacted in the U.S.,[66] initially defined "autonomous vehicle" to mean "a motor vehicle that uses artificial intelligence, sensors and global positioning system coordinates to drive itself without the active intervention of a human operator," and further defined "artificial intelligence" to mean "the use of

[63] Id. at 1078 (quoting *Recording Indus. Ass'n of Am., Inc. v. Diamond Multimedia Sys., Inc.*, 29 F. Supp. 2d 624, 630 (C.D. Cal. 1998)).

[64] Crootof & Ard (2021: 27).

[65] S. Rep. 102-294, 30 (1992).

[66] Knapp (2011).

computers and related equipment to enable a machine to duplicate or mimic the behavior of human beings."[67]

Shortly after AB511 was enacted, however, several commentators noted that the statute's definition of "autonomous vehicle" technically included vehicles that incorporated automatic collision avoidance or any of a number of other advanced driver-assistance systems common in new cars in 2011.[68] These systems used computers to temporarily control the operation of a vehicle without the intervention of the human driver, so any vehicle that incorporated them was technically subject to the onerous regulatory scheme that Nevada's legislature had intended to impose only on fully autonomous vehicles. In order to avoid effectively banning most new model cars, Nevada's legislature was forced to repeal its new law and enact a replacement that incorporated a more detailed definition of "autonomous vehicle."[69]

C. Technological Change Can Repeatedly Render a Proxy Metric Obsolete

Finally, a misspecified rule may lose its effectiveness over time as technological advances render it obsolete, necessitating repeated updates and patches to the fraying regulatory regime. Consider, for example, the export controls imposed on high performance computers in the 1990s and early 2000s. The purpose of these controls was to prevent the export of powerful computers to countries where they might be used in ways that threatened the national security of the U.S. and allied countries, such as to design missiles and nuclear weapons.[70] The government placed restrictions on the export of "supercomputers" and defined "supercomputer" in terms of the number of millions of theoretical operations per second (MTOPS) the computer could perform.[71] In 1991, "supercomputer" was defined to mean any computer capable of exceeding 195 MTOPS.[72] As the 90s progressed, however, the processing power of commercially available computers increased rapidly and companies in a number of foreign countries began to manufacture powerful computers and sell them abroad, reducing the effectiveness of U.S. export controls.[73] Restrictions that prevented U.S. companies from selling their computers globally imposed costs on the U.S. economy and harmed the international

[67] See NRS § 482A.020 (repealed 2013); NRS § 482A.30 (2011).

[68] Calo (2014).

[69] Id.

[70] See Johnston (1998); L.A. Times, *Supercomputer Export Curbs Planned* (June 8, 1991).

[71] The first export controls based on MTOPS were implemented in 1991 via a bilateral agreement between the U.S. and Japan, which at the time were the only countries in the world that manufactured computers powerful enough to be subject to the restrictions. Johnston (1998: 48). In later years, export controls were promulgated by the U.S. Department of Commerce via the Commerce Control List. See, e.g., 15 C.F.R. § 770.2 (1993).

[72] Id.; see L.A. Times, *Supercomputer Export Curbs Planned* (June 8, 1991).

[73] Steinbrecher (1995: 697).

competitiveness of the restricted companies.[74] The Clinton administration responded by raising the threshold at which export restrictions began to apply to 1500 MTOPS in 1994, to 7000 MTOPS in 1996, to 12,300 MTOPS in 1999, and three times in the year 2000 to 20,000, 28,000, and finally 85,000 MTOPS.[75]

In the late 1990s, technological advances made it possible to link large numbers of commercially available computers together into "clusters" which collectively could outperform most supercomputers.[76] At this point, it was clear to most commentators that MTOPS-based export controls were no longer effective, as high performance computers that exceeded any limit imposed could easily be produced by anyone with access to a supply of less powerful computers which would not be subject to export controls.[77] Even so, MTOPS-based export controls continued in force until 2006, when they were replaced by regulations that imposed controls based on performance in terms of Weighted TeraFLOPS, i.e., trillions of floating point operations per second.[78]

Thus, while the use of MTOPS thresholds as proxies initially resulted in well-specified export controls that effectively prevented U.S. adversaries from acquiring supercomputers, rapid technological progress repeatedly rendered the controls overinclusive and necessitated a series of amendments and revisions. The end result was a period of nearly seven years during which the existing export controls were badly misspecified due to the use of a proxy metric, MTOPS, which no longer bore any significant relation to the regime's purpose. During this period, the U.S. export control regime for high performance computers was widely considered to be ineffective and even, according to some commentators, counterproductive.[79]

IV. Mitigating Risks from Misspecification

In light of the frequency with which misspecification occurs in the regulation of emerging technology and the potential severity of its consequences, this Section suggests a few techniques for designing legal rules in such a way as to reduce the risk of misspecification and mitigate its ill effects.

The simplest way to avoid misspecification is to eschew or minimize the use of proxy terms and metrics and instead to pursue the rule's purpose directly. This is not always practicable, and when practicable it is not always desirable. "No vehicles in the park" is a better rule than "do not unreasonably annoy or endanger the safety of park visitors," in part because it reduces the cognitive burden of following, enforcing, and

[74] Id.
[75] Etter et al. (2001: 2).
[76] Picker (2001: 212).
[77] Id.; see Etter et al. (2001: 3–5).
[78] 71 FR 20876 (2006).
[79] See Picker (2001: 212).

interpreting the rule and limits the discretion of the parties charged with enforcement and interpretation, reducing the risk of decision maker error.[80] But there are nevertheless examples of successful legal rules that pursue their purposes directly. U.S. antitrust law, for example, has depended for more than a hundred years on the Sherman Antitrust Act,[81] § 1 of which simply states that any combination or contract in restraint of trade "is declared to be illegal."

Where use of a proxy is appropriate, it is often worthwhile to identify the fact that a proxy is being used to reduce the likelihood that decision makers will fall victim to Goodhart's law[82] and treat the regulation of the proxy as an end in and of itself.[83] Alternatively, the most direct way to avoid confusion regarding the underlying purpose of a rule is to simply include an explanation of the purpose in the text of the rule itself. This can be accomplished through the addition of a purpose clause (sometimes referred to as a legislative preamble or a policy statement). For example, the purpose of the Nuclear Energy Innovation and Modernization Act of 2019 is to "provide— (1) a program to develop the expertise and regulatory processes necessary to allow innovation and the commercialization of advanced nuclear reactors; (2) a revised fee recovery structure to ensure the availability of resources to meet industry needs without burdening existing licensees unfairly for inaccurate workload projections or premature existing reactor closures; and (3) a more efficient regulation of uranium recovery."

Purpose clauses can also incorporate language emphasizing that every provision of a rule should be construed in order to effectuate its purpose. This amounts to a legislatively prescribed rule of statutory interpretation, instructing courts to adopt a purposivist interpretive approach.[84] When confronted with an explicit textual command to this effect, even strict textualists are obligated to interpret a rule purposively.[85] The question of whether such an approach is generally desirable is hotly debated,[86] but in the context of AI governance, textually enforced purposivism possesses certain key advantages. The most salient of these is flexibility. The ability to flexibly update and adapt a rule in response to changes in the environment in which the rule will apply is unusually important in the regulation of emerging technologies.[87] One potential

[80] See Schauer (1988: 541).

[81] 15 U.S.C. §§ 1 *et seq*.

[82] As paraphrased by Keith Hoskins, Goodhart's law states that "every measure which becomes a target becomes a bad measure." Hoskins (1996); see Goodhart (1975) and Campbell (1979).

[83] See Manheim (2023: 3).

[84] See Rosenkranz (2002); Stack (2019).

[85] See generally Stack (2019).

[86] See, e.g., Katzmann (2014: 32); Manning (2006).

[87] See Drukarch et al. (2023), discussing the "increasing gap between the policy cycle's speed and that of technological and social change"; Marchant & Stevens (2017:252), noting that traditional administrative law processes are "not conducive to such flexible and adaptive regulatory controls" as the regulation of emerging technologies requires; Maas (2020: 247) (discussing the importance of flexible and dynamic regulatory regimes for the governance of emerging technologies).

disadvantage of mandating a purposivist interpretive approach, it should be noted, is that purposivism reallocates decision-making responsibility from Congress and executive agencies to courts, despite the fact that legislatures and agencies are, in theory, better equipped to investigate complex factual or technical problems than courts.[88] While there is little empirical evidence for or against the effectiveness of purpose clauses, they have played a key role in the legal reasoning relied on in a number of important court decisions.[89]

A regulatory regime can also require periodic efforts to evaluate whether a rule is achieving its purpose.[90] This can facilitate awareness of whether the proxy terms or metrics relied upon still correspond well to the purpose of the rule, and can provide an early warning system for misspecification. This kind of mandated periodic review of existing regulations is relatively common.[91] Existing periodic review requirements are often ineffective,[92] treated by agencies as box-checking activities rather than genuine opportunities for careful retrospective analysis of the effects of regulations.[93] However, many experts continue to recommend well-implemented retrospective review requirements as an effective tool for improving policy decisions.[94] The Administrative Conference of the United States has repeatedly pushed for increased use of retrospective review, as has the internationally-focused Organization for Economic Co-Operation and Development (OECD).[95] Additionally, retrospective review of regulations often works well in countries outside of the U.S.[96] As the examples in Sections II and III demonstrate,

[88] See, e.g., *Turner Broad. Sys., Inc. v. FCC*, 520 U.S. 180, 195 (1997) ("We owe Congress' findings deference in part because the institution is far better equipped than the judiciary to amass and evaluate the vast amounts of data bearing upon legislative questions."); *Chevron, U.S.A., Inc. v. Nat. Res. Def. Council, Inc.*, 467 U.S. 837, 865 (1984) (deferring to administrative agency technical expertise in on a question of statutory interpretation because "[j]udges are not experts in the field").

[89] See, e.g., *King v. Burwell*, 576 U.S. 473, 482 (2015) (upholding the eligibility of persons in states with federally created health exchanges for premium tax credits under the Affordable Care Act on the basis that a court "cannot interpret federal statutes to negate their own stated purposes"); *Merrill Lynch, Pierce, Fenner & Smith Inc. v. Dabit*, 547 U.S. 71, 86 (2006) ("A narrow reading of the statute would undercut the effectiveness of the 1995 Reform Act and thus run contrary to SLUSA's stated purpose."); *Alabama Dep't of Revenue v. CSX Transp., Inc.*, 575 U.S. 21, 28 (2015).

[90] See generally Bennear & Wiener (2021).

[91] The first instance of this kind of periodic ex-post regulatory impact analysis appeared in Executive Order 12,044 (1970), which mandated that agencies "periodically review their existing regulations to determine whether they are achieving the policy goals of this Order." Improving Government Regulations, 43 FR 12661. A similar periodic review requirement is present in § 610 of the Regulatory Flexibility Act, 5 U.S.C. § 610(a), which requires regulators to periodically review existing rules and consider eliminating those which are duplicative, unduly burdensome, or unnecessary.

[92] Commentators have complained that inconsistent compliance with § 610 by agencies has limited the effectiveness of that provision. See See (2006). The effectiveness of ex post regulatory impact analysis generally, as currently implemented in the U.S., has also been criticized as "ad hoc and largely unmanaged" and "patchy and unsystematic." See Wiener & Ribeiro (2016: 6).

[93] Bennear & Wiener (2021: 15).

[94] See Sunstein (2014); Bennear & Wiener (2021).

[95] Bennear & Wiener (2021); OECD (2012).

[96] Coglianese (2013: 12–13).

one consistent theme in case studies of the governance misspecification problem is that rules tend to become misspecified over time as the regulated technology evolves. The Outer Space Treaty of 1967, § 1201(a)(2) of the DMCA, and the Clinton Administration's supercomputer export controls were all well-specified and effective when implemented, but each measure became ineffective or counterproductive soon after being implemented because the proxies relied upon became obsolete. Ideally, rulemaking would move at the pace of technological improvement, but there are a number of institutional and structural barriers to this sort of rapid updating of regulations. Notably, the Administrative Procedure Act requires a lengthy "notice and comment" process for rulemaking and a 30-day waiting period after publication of a regulation in the Federal Register before the regulation can go into effect.[97] There are ways to waive or avoid these requirements, including regulating via the issuance of nonbinding guidance documents rather than binding rules,[98] issuing an immediately effective "interim final rule" and then satisfying the APA's requirements at a later time,[99] waiving the publication or notice and comment requirements for "good cause,"[100] or legislatively imposing regulatory deadlines.[101] Many of these workarounds are limited in their scope or effectiveness, or vulnerable to legal challenges if pursued too ambitiously, but finding some way to update a regulatory regime quickly is critical to mitigating the damage caused by misspecification.[102]

There is reason to believe that some agencies, recognizing the importance of AI safety to national security, will be willing to rapidly update regulations despite the legal and procedural difficulties. Consider the Commerce Department's recent response to repeated attempts by semiconductor companies to design chips for the Chinese market that comply with U.S. export control regulations while still providing significant utility to purchasers in China looking to train advanced AI models. After Commerce initially imposed a license requirement on the export of advanced AI-relevant chips to China in October 2022, Nvidia modified its market-leading A100 and H100 chips to comply with the regulations and proceeded to sell the modified A800 and H800 chips in China.[103] On October 17, 2023, the Commerce Department's Bureau of Industry and Security announced a new interim final rule that would prohibit the sale of A800 and H800 chips in China and waived the normal 30-day waiting period so that the rule became effective less than a week after it was announced.[104] Commerce Secretary Gina Raimondo stated

[97] See 5 U.S.C. § 553.
[98] See Cortez (2014).
[99] See Asimow (1999); Hickman & Thomson (2016); Scherber (2014).
[100] See Schneider, K. (2021).
[101] See Gersen & O'Connell (2008).
[102] See Arnold & Van Arsdale (2023).
[103] Edwards, B. (2023).
[104] NVIDIA Corporation, SEC Filing (2023).

publicly that "'[i]f [semiconductor companies] redesign a chip around a particular cut line that enables them to do AI, I'm going to control it the very next day."[105]

V. The Governance Misspecification Problem and Artificial Intelligence

While the framework of governance misspecification is applicable to a wide range of policy measures, it is particularly well-suited to describing issues that arise regarding legal rules governing emerging technologies. H.L.A. Hart's prohibition on "vehicles in the park" could conceivably have been framed by an incautious drafter who did not anticipate that using "vehicle" instead of some more detailed proxy term would create ambiguity. Avoiding this kind of misspecification is simply a matter of careful drafting. Suppose, however, that the rule was formulated at a point in time when "vehicle" was an appropriate proxy for a well-understood category of object, and the rule later became misspecified as new potential vehicles that had not been conceived of when the rule was drafted were introduced. A rule drafted at a historical moment when all vehicles move on either land or water is unlikely to adequately account for the issues created by airplanes or flying drones.[106]

In other words, rules created to govern emerging technologies are especially prone to misspecification because they are created in the face of a high degree of uncertainty regarding the nature of the subject matter to be regulated, and rulemaking under uncertainty is difficult.[107] Furthermore, as the case studies discussed in Sections II and III show, the nature of this difficulty is such that it tends to result in misspecification. For instance, misspecification will usually result when an overconfident rulemaker makes a specific and incorrect prediction about the future and issues an underinclusive rule based on that prediction. This was the case when Congress addressed the AHRA exclusively to digital audio tape recorders and ignored computers. Rules created by rulemakers who want to regulate a certain technology but have only a vague and uncertain understanding of the purpose they are pursuing are also likely to be misspecified.[108] Hence the CFAA, which essentially prohibited "doing bad things with a computer," with disastrous results.

The uncertainties associated with emerging technologies and the associated risk of misspecification increase when the regulated technology is poorly understood. Rulemakers may simply overlook something about the chosen proxy due to a lack of understanding of the proxy or the underlying technology, or due to a lack of experience drafting the kinds of regulations required. The first-of-its-kind Nevada law intended to

[105] Martin (2023).

[106] See *McBoyle v. United States*, 283 U.S. 25 (1931).

[107] See Drukarch et al. (2023); Johnson (2022).

[108] Cf. Crootof & Ard (2021: 62–63), discussing the issues caused by laws "based on underexplored intuitions regarding the ideal default."

regulate fully autonomous vehicles that accidentally regulated a broad range of features common in many new cars is an example of this phenomenon. So is the DMCA provision that was intended to regulate "black box" devices but, by its terms, also applied to raw computer code.

If the difficulty of making well-specified rules to govern emerging technologies increases when the technology is fast-developing and poorly understood, advanced AI systems are something of a perfect storm for misspecification problems. Cutting-edge deep learning AI systems differ from other emerging technologies in that their workings are poorly understood, not just by legislators and the public, but by their creators.[109] Their capabilities are an emergent property of the interaction between their architecture and the vast datasets on which they are trained. Moreover, the opacity of these models is arguably different in kind from the unsolved problems associated with past technological breakthroughs, because the models may be fundamentally uninterpretable rather than merely difficult to understand.[110] Under these circumstances, defining an ideal specification in very general terms may be simple enough, but designing legal rules to operationalize any such specification will require extensive reliance on rough proxies. This is fertile ground for misspecification.

There are a few key proxy terms that recur often in existing AI governance proposals and regulations. For example, a number of policy proposals have suggested that regulations should focus on "frontier" AI models.[111] When Google, Anthropic, OpenAI, and Microsoft created an industry-led initiative to promote AI safety, they named it the Frontier Model Forum.[112] Sam Altman, the CEO of OpenAI, has expressed support for regulating "frontier systems."[113] The government of the U.K. has established a "Frontier AI Taskforce" dedicated to evaluating risks "at the frontier of AI."[114]

In each of these proposals, the word "frontier" is a proxy term that stands for something like "highly capable foundation models that could possess dangerous capabilities sufficient to pose severe risks to public safety."[115] Any legislation or regulation that relied on the term "frontier" would also likely include a statutory definition of the word,[116] but as several of the historical examples discussed in Sections II and III showed, statutory definitions can themselves incorporate proxies that result in

[109] See Knight (2017); Zhang et al. (2021).
[110] Wang et al. (2023).
[111] See Anderljung et al. (2023); Toner & Fist (2023); Toner et al. (2023).
[112] OpenAI (2023).
[113] https://twitter.com/sama/status/1720165289864712541.
[114] Frontier AI Taskforce (2023).
[115] Anderljung et al. (2023: 2).
[116] Cf. EO 14110 (October 30, 2023: § 3(k)) (defining "dual-use foundation model" with a detailed provision that begins "an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters…").

misspecification. The above definition, for instance, may be underinclusive because some models that cannot be classified as "highly capable" or as "foundation models" might also pose severe risks to public safety.

The most significant AI-related policy measure that has been issued in the U.S. to date is Executive Order (EO) 14110 on the "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence."[117] Among many other provisions, the EO imposes reporting requirements on certain AI models and directs the Department of Commerce to define the category of models to which the reporting requirements will apply.[118] Prior to the issuance of Commerce's definition, the EO provides that the reporting requirements apply to models "trained using a quantity of computing power greater than 10^{26} integer or floating-point operations, or using primarily biological sequence data and using a quantity of computing power greater than 10^{23} integer or floating-point operations," as well as certain computing clusters.[119] In other words, the EO uses operations as a proxy metric for determining which AI systems are sufficiently capable and/or dangerous that they should be regulated.[120] This kind of metric, which is based on the amount of computing power used to train a model, is known as a "compute threshold" in the AI governance literature.[121]

A proxy metric such as an operations-based compute threshold is almost certainly necessary to the operationalization of the EO's regulatory scheme for governing frontier models.[122] Even so, the example of the U.S. government's ultimately ineffective and possibly counterproductive attempts to regulate exports of high performance computers using MTOPS is a cautionary tale about how quickly a compute-based proxy can be rendered obsolete by technological progress. The price of computing resources has, historically, fallen rapidly, with the amount of compute available for a given sum of money doubling approximately every two years as predicted by Moore's Law.[123] Additionally, because of improvements in algorithmic efficiency, the amount of compute required to train a model to a given level of performance has historically decreased over time as well.[124] Because of these two factors, the cost of training AI models to a given level of capability has fallen precipitously over time; for instance, between 2017 and

[117] EO 14110 (October 30, 2023).

[118] See id. at § 4.2(b).

[119] Id. at § 4.2(b)(1).

[120] Recall that the Department of Commerce used a related metric (floating-point operations per second, or FLOPS) to replace the obsolete MTOPS metric for determining which computers were capable and/or dangerous enough to be subjected to export controls.

[121] See, e.g., Egan & Heim (2023: 5); Smith et al. (2023: 4).

[122] Note that descriptions of AI governance measures will frequently use nested systems of proxy terms. Thus, a compute threshold is a proxy metric used to define "frontier model," which is itself a proxy term.

[123]Gordon Moore initially predicted that the number of transistors on an integrated circuit would double every year, but later revised this prediction to one doubling every two years. Compare Moore (1965), with Moore (1975: 11–13).

[124] Pilz et al. (2023).

2021, the cost of training a rudimentary model to classify images correctly with 93% accuracy on the image database ImageNet fell from $1000 to $5.[125] This phenomenon presents a dilemma for regulators: the cost of acquiring computational resources exceeding a given threshold will generally decrease over time even as the capabilities of models trained on a below-threshold amount of compute rises. In other words, any well-specified legal rule that uses a compute threshold is likely to be rendered both overinclusive and underinclusive soon after being implemented.

Export controls intended to prevent the proliferation of the advanced chips used to train frontier AI models face a similar problem. Like the Clinton Administration's supercomputer export controls, the Biden administration's export controls on chips like the Nvidia A800 and H800 are likely to become misspecified over time. As algorithmic efficiency increases and powerful chips become cheaper and easier to acquire, existing semiconductor export controls will gradually become both overinclusive (because they pointlessly prohibit the export of chips that are already freely available overseas) and underinclusive (because powerful AI models can be trained using chips not covered by the export controls).

The question of precisely how society should respond to these developments over time is beyond the scope of this paper. However, to delay the onset of misspecification and mitigate its effects, policymakers setting legal rules for AI governance should consider the recommendations outlined in Section IV, above. So, the specifications for export controls on semiconductors—proxies for something like "chips that can be used to create dangerously powerful AI models"—should be updated quickly and frequently as needed, to prevent them from becoming ineffective or counterproductive. The Bureau of Industry and Security has already shown some willingness to pursue this kind of frequent, flexible updating.[126] More generally, given the particular salience of the governance misspecification problem to AI governance, legislators should consider mandating particularly frequent review of the effectiveness of important AI regulations and empowering administrative agencies to update regulations rapidly as necessary. Laws or regulations setting compute thresholds that are likely to be the subject of litigation should incorporate clear purpose statements articulating the ulterior purpose behind the use of a compute threshold as a proxy, and should be interpreted consistently with those statements. And where it is possible to eschew the use of proxies without compromising the enforceability or effectiveness of a rule, legislators and regulators should consider doing so.

[125] Id.

[126] See Martin (2023) and the discussion of Commerce Secretary Raimondo's comments in Section IV above.

VI. Conclusion

This article has attempted to elucidate a newly developed concept in governance, i.e., the problem of governance misspecification. In presenting this concept along with empirical insights from representative case studies, we hope to inform contemporary debates around AI governance by demonstrating one common and impactful way in which legal rules can fail to effect their purposes. By framing this problem in terms of "misspecification," a concept borrowed from the technical AI safety literature, this article aims both to introduce valuable insights from that field to scholars of legal philosophy and public policy and to introduce technical researchers to some of the more practically salient legal-philosophical and governance-related challenges involved in AI legislation and regulation. Additionally, we have offered a few specific suggestions for avoiding or mitigating the harms of misspecification in the AI governance context, namely eschewing the use of proxy terms or metrics where feasible, clear statements of statutory purpose, and flexibly applied, rapidly updating, periodically reviewed regulations.

A great deal of conceptual and empirical work remains to be done regarding the nature and effects of the governance misspecification problem and best practices for avoiding and responding to it. For instance, this article does not contain any in-depth comparison of the incidence and seriousness of misspecification outside of the context of rules governing emerging technologies. Additionally, empirical research analyzing whether and how purpose clauses and similar provisions can effectively further the purposes of legal rules would be of significant practical value.

References

Altman, S. (2023). https://twitter.com/sama/status/1720165289864712541.

Anderljung, M. et al. (2023). Frontier AI Regulation: Managing Emerging Risks to Public Safety. https://arxiv.org/abs/2307.03718.

Arnold, M. & Van Arsdale, S. (2023). Rapidly and/or Frequently Updating Regulations - Legal Mechanisms and Barriers. Unpublished Memorandum on File with Authors.

Asimow, M. (1999). Interim-Final Rules: Making Haste Slowly. *Administrative Law Review*, 51(3), 703–755.

Bas, G. (2023). Operationalising the definition of highly capable AI. https://static1.squarespace.com/static/60c0fe48b1480d2dddf3bff9/t/654befde71b69c635988fb9e/1699475423076/Operationalising+the+definition+of+highly+capable+AI.pdf,

Bennear, L., & Wiener, J. (2021). Periodic Review of Agency Regulation. Report for the Administrative Conference of the United States. https://www.acus.gov/sites/default/files/documents/ACUS%20-%20Periodic%20Review%20-%20Periodic%20Review%20of%20Agency%20Regulation%202021%2006%2007%20final%20%281%29.pdf

Bricken, T. et al. (2023). Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. https://transformer-circuits.pub/2023/monosemantic-features/index.html

Calo, R. (2014). The Case for a Federal Robotics Commission. The Brookings Institution. Retrieved December 5, 2023, from https://www.brookings.edu/articles/the-case-for-a-federal-robotics-commission/.

Campbell, D.T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*, 2(1), 67–90.

Chung, C. (2010). The Computer Fraud and Abuse Act: How Computer Science Can Help with the Problem of Overbreadth. *Harvard Journal of Law and Technology*, 24, 233–256.

Coglianese, C. (2013). Thinking Ahead, Looking Back: Assessing the Value of Regulatory Impact Analysis and Procedures for Its Use. *Korean Journal of Law and Legislation*, 3(1), 5–27.

Computer Crime and Intellectual Property Section, U.S. Department of Justice (1998). The National Information Infrastructure Protection Act of 1996. Retrieved November 30, 2023, from https://www.hsdl.org/?view&did=439224.

Cortez, N. (2014). Regulating Disruptive Innovation. *Berkeley Technology Law Journal*, 29, 175–228.

Curtiss, T. (2016). Computer Fraud and Abuse Act Enforcement: Cruel, Unusual, and Due for Reform Due for Reform. *Washington Law Review*, 91(4), 1813–1850.

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. https://arxiv.org/abs/1702.08608.

Drukarch, H. et al. (2023). An iterative regulatory process for robot governance. *Data & Policy*, 5:e-8.

Easterbrook, F. (1983). Statutes' Domains. *University of Chicago Law Review*, 50, 533–552.

Easterbrook, F. (1990). What Does Legislative History Tell Us? *Chicago-Kent Law Review*, 66, 441–450.

Edwards, B. (2023). US surprises Nvidia by speeding up new AI chip export ban. https://arstechnica.com/information-technology/2023/10/ai-chip-wars-us-curbs-nvidia-gpu-chip-exports-sooner-than-expected/

Eskridge, W. (1990). The New Textualism. *UCLA Law Review*, 37, 621—691.

Executive Order 14410, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence." October 30, 2023. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

Etter, D. et al. (2001). Export Control of High Performance Computing: Analysis and Alternative Strategies. Defense Science and Technology Reports. https://apps.dtic.mil/sti/tr/pdf/ADA397730.pdf.

Frontier AI Taskforce (2023). Frontier AI Taskforce: first progress report. https://www.gov.uk/government/publications/frontier-ai-taskforce-first-progress-report/frontier-ai-taskforce-first-progress-report.

Freitas-Groff, Z. (2023). Persistence in Policy: Evidence from Close Votes. https://zachfreitasgroff.b-cdn.net/FreitasGroff_Policy_Persistence.pdf.

Fuller, L. (1956). Human Purpose and Natural Law. *Journal of Philosophy*, 53(22), 697–705.

Fuller, L. (1958). Positivism and Fidelity to Law—A Reply To Professor Hart. *Harvard Law Review*, 71, 630–672.

Garthoff, R. (1980). Banning the Bomb in Outer Space. *International Security*, 5(3), 25–40.

Gersen, J. & O'Connell, A. (2008). Deadlines in Administrative Law, *University of Pennsylvania Law Review*, 156, 923–990.

Gigerenzer, G., & Engel, C. (2006). Law and heuristics: An interdisciplinary venture. In Gigerenzer, G., & Engel, C. *Heuristics and the law*. Cambridge: The MIT Press, 1–16.

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In Gigerenzer, G., Todd, P. M., & the ABC Research Group (Eds.) *Simple heuristics that make us smart*. Oxford: Oxford University Press, 3–34.

Ginsburg, J. (1999). Copyright Legislation for the "Digital Millennium." *Columbia-VLA Journal of Law and the Arts*, 23, 137–179.

Goodhart, C.A.E. (1975). Problems of monetary management: the UK experience. In *Papers in Monetary Economics* (Reserve Bank of Australia).

Hadfield-Menell, D., & Hadfield, G. (2018). Incomplete Contracting and AI Alignment. https://arxiv.org/abs/1804.04268.

Hart, H.L.A. (1958). Positivism and the Separation of Law and Morals. *Harvard Law Review*, 71, 593–629.

Hart, H., & Sacks, A. (1995). *The Legal Process: Basic Problems in the Making and Application of Law*. Foundation Press, Eskridge, N., & Frickey, P., eds.

Hickman, K., & Thomson, M. (2016). Open Minds and Harmless Errors: Judicial Review of Postpromulgation Notice and Comment. *Cornell Law Review*, 101, 261–324.

Hoskins, K. (1996). The 'awful idea of accountability': inscribing people into the measurement of objects. In Munro, R. & Mouritsen, J. (Eds.) *Accountability: Power, ethos and the technologies of managing.* 265.

Howlett, M. (2009). Governance modes, policy regimes and operational plans: A multi-level nested model of policy instrument choice and policy design. *Policy Sciences*, 42(1), 73–89.

Johnson, W. (2022). Flexible regulation for dynamic products? The case of applying principles-based regulation to medical products using artificial intelligence. *Law, Innovation and Technology*, 14(2), 205-236.

Johnston, R. (1998). U.S. Export Control Policy in the High Performance Computer Sector. *The Nonproliferation Review*, Winter 1998, 44–59.

Katzmann, R. (2012). Statutes, *New York University Law Review*, 87, 637–667.

Katzmann, R. (2014). *Judging Statutes*. Oxford University Press.

Kerr, O. (2003). Cybercrime's Scope: Interpreting "Access" and "Authorization" in Computer Misuse Statutes. *New York University Law Review*, 78, 1596–1668.

Kerr, O. (2010). Vagueness Challenges to the Computer Fraud and Abuse Act, *Minnesota Law Review*, 94, 1561–1587.

Knapp, A. (2011). Nevada Passes Law Authorizing Driverless Cars. *Forbes*.

Knight, W. (2017). The Dark Secret at the Heart of AI. *MIT Technology Review*, April 11, 2017. https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/.

Heim, L. & Egan, J. (2023). Oversight for Frontier AI through a Know-Your-Customer Scheme for Compute Providers. https://arxiv.org/abs/2310.13625.

Howlett, M., & Rayner, J. (2018). Coherence, congruence and consistency in policy mixes. In M. Howlett & I. Mukherjee (Eds.), *Routledge handbook of policy design* (pp. 389–403). Routledge.

Maas, M. (2020). *Artificial Intelligence Governance under Change: Foundations, Facets, Frameworks*.

Maggs, G. (1992). Reducing the Costs of Statutory Ambiguity: Alternative Approaches and the Federal Courts Study Committee, *Harvard Journal on Legislation*, 29, 123–173.

Manheim, D. (2023). Building less-flawed metrics: Understanding and creating better measurement and incentive systems. *Patterns*, 4(10).

Manning, J. (2006). What Divides Textualists from Purposivists? *Columbia Law Review*, 106, 70–111.

Manning, J. (2011). The New Purposivism. *Supreme Court Review*, 2011, 113–182.

Marchant, G. & Stevens, Y. (2017). Resilience: A New Tool in the Risk Governance Toolbox for Emerging Technologies. *UC Davis Law Review*, 51, 233–271.

Martin, P. (December 2, 2023). 'We cannot let China get these chips': Commerce Secretary Raimondo says more funding needed for AI export controls. Fortune, December 2, 2023. https://fortune.com/2023/12/02/ai-chip-export-controls-china-nvidia-raimondo/.

Martínez, E., & Tobia, K. (2023). What Do Law Professors Believe about Law and the Legal Academy? *Georgetown Law Journal*, 112, 111–189.

Moore, G. (1965). Cramming More Components onto Integrated Circuits. *Electronics*, 38, 114–117.

Moore, G. (1975). Progress in Digital Integrated Electronics. *Technical Digest*, *International Electronic Devices Meeting, IEEE*, 11–13.

Nimmer, D. (2004). Codifying Copyright Comprehensively. *UCLA Law Review*, 51, 1233–1387.

NVIDIA Corporation, October 24, 2023, SEC Form 8-K filing. https://www.sec.gov/Archives/edgar/data/1045810/000104581023000221/nvda-20231023.htm.

OpenAI (2016). Faulty reward functions in the wild. https://openai.com/research/faulty-reward-functions.

OpenAI (2023). Frontier Model Forum. https://openai.com/blog/frontier-model-forum.

Organization for Economic Co-operation and Development (2012). Recommendation of the Council on Regulatory Policy Governance.

Ortega, P. et al. (2018). Building safe artificial intelligence: specification, robustness, and assurance. https://deepmindsafetyresearch.medium.com/building-safe-artificial-intelligence-52f5f75058f1

Picker, C. (2001). The View from 40,000 Feet: International Law and the Invisible Hand of Technology. *Cardozo Law Review*, 23(1), 149–219.

Pickett, J. (2019). Public Opinion and Criminal Justice Policy: Theory and Research. *Annual Review of Criminology*, 2, 405–428.

Pilz, K. et al. (2023). Increased Compute Efficiency and the Diffusion of AI Capabilities. https://arxiv.org/pdf/2311.15377.pdf.

Radin, M. (1930). Statutory Interpretation. *Harvard Law Review*, 43, 863–885.

Rosenkranz, N. (2002). Federal Rules of Statutory Interpretation. *Harvard Law Review*, 115, 2085–2157.

Rudin, C. et al. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16, 1–85.

Schauer, F. (1988). Formalism. *Yale Law Journal*, 97, 509–548.

Schauer, F. (1991). *Playing By the Rules: A Philosophical Examination of Rule-Based Decision-Making in Law and in Life.* Oxford University Press.

Scherber, L. (2014). Interim Final Rules Over Time: A Brief Empirical Analysis, GW Regulatory Studies Center. https://regulatorystudies.columbian.gwu.edu/interim-final-rules-over-time-brief-empirical-analysis.

Schneider, K. (2021). Judicial Review of Good Cause Determinations Under the Administrative Procedure Act. *Stanford Law Review*, 73, 237–283.

See, M. (2006). Willful Blindness: Federal Agencies' Failure to Comply with the Regulatory Flexibility Act's Periodic Review Requirement and Current Proposals to Invigorate the Act. *Fordham Urban Law Journal*, 33, 1199–1255.

Smith, G. et al. (2023). Industry and Government Collaboration on Security Guardrails for AI Systems. RAND Corporation Conference Proceedings. https://www.rand.org/content/dam/rand/pubs/conf_proceedings/CFA2900/CFA2949-1/RAND_CFA2949-1.pdf.

Stack, K. (2019). The Enacted Purposes Canon. *Iowa Law Review*, 105, 283–339.

Steinbrecher, B. (1995). The Impact of the Clinton Administration's Export Promotion Plan on U.S. Exports of Computers and High-Technology Equipment. *University of Pennsylvania Journal of International Business Law*, 15, 675–706.

Struchiner, N. et al. (2020). An experimental guide to vehicles in the park. *Judgment and Decision Making*, 15(3), 312–329.

Sunstein, C. (2014). The Regulatory Lookback. *Boston University Law Review*, 94, 579–602.

Toner, H. & Fist, T. (2023). Regulating the AI Frontier: Design Choices and Constraints. Center for Security and Emerging Technology. https://cset.georgetown.edu/article/regulating-the-ai-frontier-design-choices-and-constraints/.

Toner, H. et al. (2023). Skating to Where the Puck Is Going: Anticipating and Managing Risks from Frontier AI Systems. Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/skating-to-where-the-puck-is-going.

Van Geet, M. et al. (2021). The importance of policy design fit for effectiveness: a qualitative comparative analysis of policy integration in regional transport planning. *Policy Sciences*, 54, 629–662.

Vogel, D. (2008). When Consumers Oppose Consumer Protection: The Politics of Regulatory Backlash. *Journal of Public Policy*, 10(4), 449–470.

Wang, T. et al. (2023). Forbidden Facts: An Investigation of Competing Objectives in Llama-2. https://arxiv.org/abs/2312.08793.

Wiener, J., & Ribeiro, D. (2016). Environmental Regulation Going Retro: Learning Foresight From Hindsight. *Journal of Land Use*, 32(1), 1–73.

Wittgenstein, L. (1953). *Philosophical Investigations*. Macmillan Publishers.

Wu, T. (2013). Fixing the Worst Law in Technology. *The New Yorker*.

Zhang, Y. et al. (2021). A Survey on Neural Network Interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5), 726–742.